**FINDING THE BEST LOCATION FOR A MEXICAN RESTAURANT IN MADRID**

# C:\Users\jnascime\Desktop\transferir.jpg

**Done by: João Nascimento**

# **Problem**

I'll be trying to find the best location for a mexican restaurant in Madrid, Spain. To do this, i'll use machine learning techniques (clustering) and data analysis (some visualization techniques). During the analysis phase, the data will be transformed to make it easy for the model to process it. After that, a modeling process will be carried out, and this statistical analysis will give us the best places to have a mexican restaurant in Madrid. I love mexican food and the city of Madrid as well, so it's a combination of both worlds for me. Besides, there are many mexicans living in Spain, and as such, finding the best location to build a mexican restaurant in the capital of the country is of most interest.

# **Data Presentation**

The data comes from two websites: The Madrid City Hall's Web Portal: The data contains updated information about the population per country and per nationality in the city of Madrid (in excel format), and it allows us to find the best location of a restaurant based on the customer's nationality (I will assume that people will be more attracted to a place that matches the culture of their own countries). The Foursquare Api: This data will be accessed via Python, and used to obtain the most common venues per neighborhood in the city of Madrid, allowing us to know how are the city's venues distributed.

# **Methodology**

We will use statistical exploration and data visualizations. Regarding machine learning, we will use the clustering method (K-Means algorithm). To solve these location problems, we need consumer’s data, but our focus was on the population’s nationality (it will be assumed that people will be more attracted to a place that matches the culture of their own countries, as in the end, it is also about having a piece of your own country).

Here is an example of the data used:

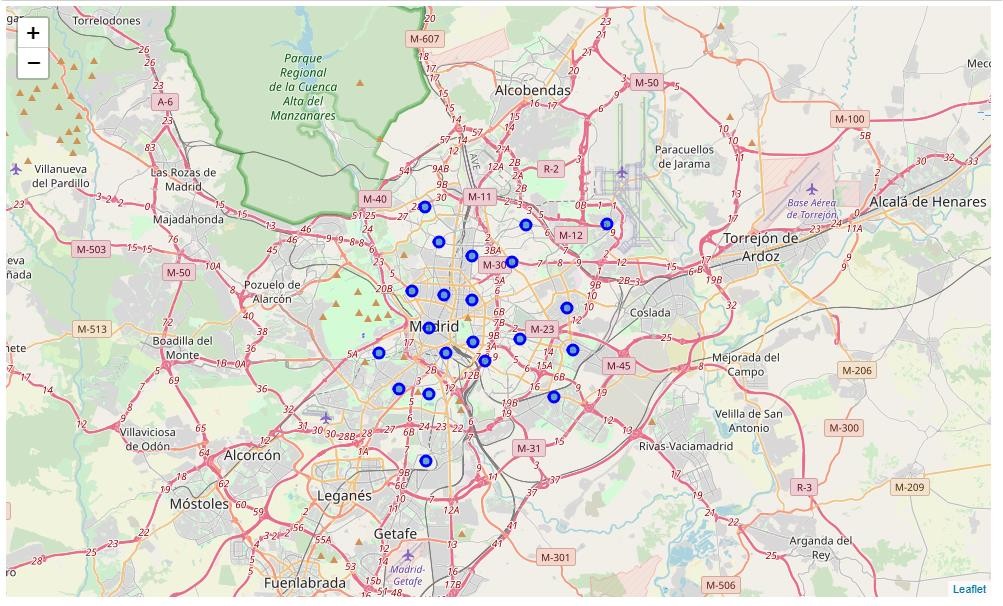
|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Country Total** | | **Centro** |  | **Arganzuela** | **Retiro** |  | **Salamanca** | **Chamartin** | **Tetuán** |
| Rumanía | 450360 | 8150 | | 7540 | 4800 | | 7530 | 6800 | 14680 |
| China | 372760 | 15080 | | 13560 | 5640 | | 7550 | 6520 | 19880 |
| Ecuador | 239530 | 6470 | | 7410 | 2650 | | 6190 | 3800 | 13950 |
| Venezuela | 233590 | 15630 | | 9130 | 6380 | | 15640 | 9330 | 13100 |
| Colombia | 226180 | 9980 | | 7170 | 4830 | | 8030 | 5510 | 8220 |
| Marruecos | 219090 | 11010 | | 3900 | 1840 | | 3220 | 2800 | 13930 |
| Italia | 203080 | 30300 | | 12190 | 8400 | | 18170 | 10600 | 11940 |
| Perú | 188290 | 5630 | | 5210 | 2530 | | 6120 | 4190 | 9650 |
| Paraguay | 186820 | 3640 | | 4740 | 2370 | | 5210 | 6570 | 33110 |
| República Do | 175110 | 3650 | | 6540 | 2040 | | 3440 | 3220 | 22720 |
| Honduras | 159810 | 1490 | | 2280 | 2320 | | 3320 | 3370 | 7550 |

This data contains information about the immigrant population of Madrid’s neighborhoods.

The Foursquare API was used to obtain the data about the venues in each neighborhood This is an example of the transformed data:

|  |
| --- |
| **Neighborhoo Latitude Longitude** |
| Centro 40415347 -3707371 |
| Arganzuela 40402733 -3695403 |
| Retiro 40408072 -3676729 |
| Salamanca 4043 -3677778 |
| Chamartin 40453333 -36775 |
| Tetuán 40460556 -37 |
| Chamberí 40432792 -3697186 |
| Fuencarral-El 40478611 -3709722 |
| Moncloa-Arav 40435151 -3718765 |
| Latina 40402461 -3741294 |
| Carabanchel 40383669 -3727989 |
| Usera 40381336 -3706856 |
| Puente de Val 40398204 -3669059 |
| Moratalaz 40409869 -3644436 |
| Ciudad Lineal 4045 -365 |

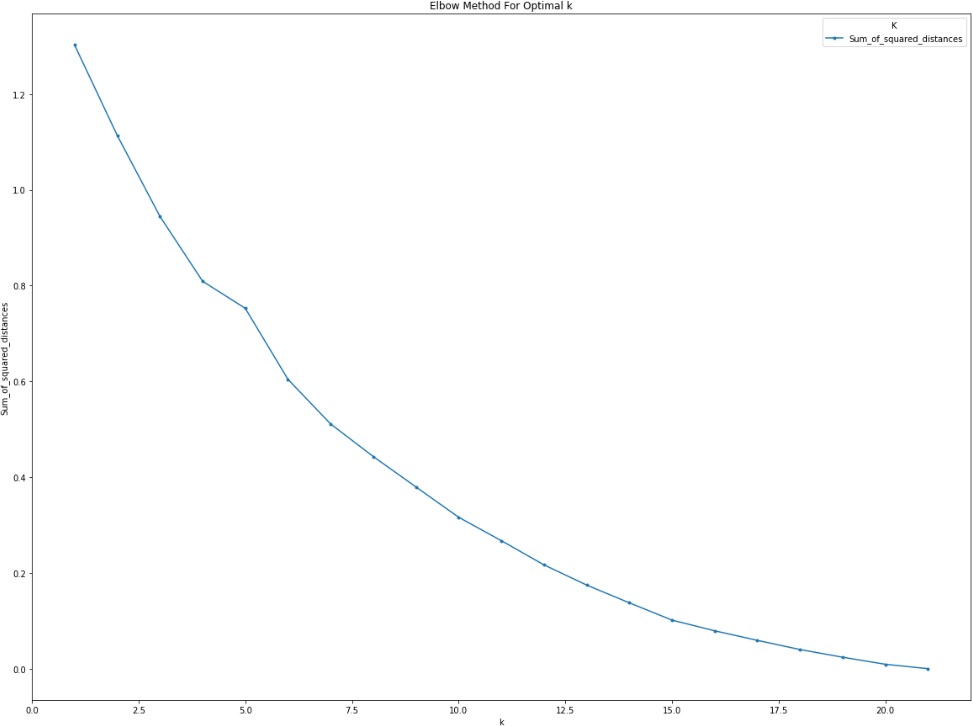
The neighborhoods were then plotted into a map of Madrid.



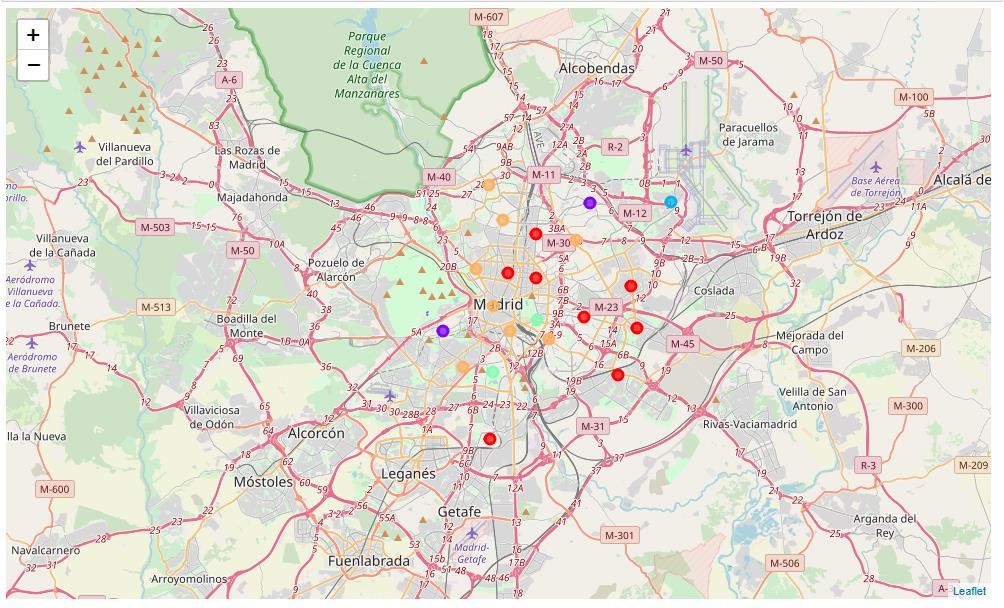
The next step was to obtain the nearby venues by neighborhood with theh their respective coordinates:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Neighborhoo Neighborhoo Neighborhoo Venue Venue LatitudVenue Longit Venue Catego** | | | | | | |
| Centro | 40415347 | -3707371 | Plaza Mayor | 4,0415E+16 | -3,7076E+16 | Plaza |
| Centro | 40415347 | -3707371 | Mercado de S | 4,0415E+15 | -3,709E+16 | Market |
| Centro | 40415347 | -3707371 | La Taberna de | 4,0415E+16 | -3,7081E+15 | Other Nightlif |
| Centro | 40415347 | -3707371 | The Hat Madr | 4,0414E+16 | -3,7071E+14 | Hotel |
| Centro | 40415347 | -3707371 | Amorino | 4,0416E+15 | -3,7084E+16 | Ice Cream Sho |
| Centro | 40415347 | -3707371 | Botao | 4,0414E+15 | -3,7081E+15 | Spanish Resta |
| Centro | 40415347 | -3707371 | Chocolate | 4,0417E+16 | -3,7068E+16 | Chocolate Sh |
| Centro | 40415347 | -3707371 | Pinkleton & W | 4,0415E+15 | -3,7091E+16 | Wine Bar |

The next step is the segmentation, but first we need to define the appropriate number of clusters. (elbow method - consists in plotting a hypothetical and usually large number of clusters in our data, and draw a curve representing the squared distances between each cluster.



The distances start reducing from cluster 5 on, so it was determined that the optimal number of clusters for this problem was 5.



Now it is possible to examine the data of each cluster:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Barajas** |  | **Neighborhoo** | **Cluster Labels** |  | **1st Most Common Venue** | **2nd Most Common Venue** |
| 3140 | | Centro | 0 | | Spanish Restaurant | Tapas Restaurant |
| 740 Villa de Vallec | | | 0 Food | | | Spanish Restaurant |
| 1910 Retiro | | | 0 Spanish Restaurant | | | Supermarket |
| 3370 Ciudad Lineal | | | 0 Spanish Restaurant | | | Burger Joint |
| 570 VicÃ¡lvaro | | | 0 Spanish Restaurant | | | Breakfast Spot |
| 2580 Chamartin | | | 0 Spanish Restaurant | | | Restaurant |
| 920 Usera | | | 0 Seafood Restaurant | | | Bubble Tea Shop |
| 910 TetuÃ¡n | | | 0 Spanish Restaurant | | | Brazilian Restaurant |

# **Results**

* **Cluster One:**

Mostly inhabited by south Americans, Europeans, and north Americans. The most common venues are tapas restaurants and Argentinian restaurants, among many others.

* **Cluster Two:**

This cluster is composed only by 2 different population kinds: Ukrainian and Dominican Republic. The most common venues are gyms and Asian restaurants.

* **Cluster Three:**

This cluster is only composed by Bangladeshi people. The most common places are falafel restaurants and fish markets

* **Cluster four:**

Again, only people from two countries seems to live in this clusters. Ecuador and Bolivia. The most common venues are nightclubs and soccer fields.

* **Cluster Five:**

Some of the main countries here are Rumania, France and Honduras. The most common venues are Mexican restaurants, Chinese restaurants and breweries.

# **Discussion**

Each cluster has its own characteristics and also common things with other clusters. In a study of this size, to make good predictions more data is needed (socio-demographic data - income level, number of children, education level, etc).

# 

# **Conclusions**

If we want to open a new Mexican restaurant in Madrid, we need to study the population in that area. If we exam cluster 1, we notice that the population are mostly Latinos, (South American countries), and Latin restaurants can be found, like Argentinian restaurants, tapas restaurants, as well as Italian restaurants. On the other hand, clusters 4 and 5 would make a good match. Looking at the venues in these clusters, it is possible to find one Mexican restaurant, and a good bunch of fast food, Argentinian, and south American restaurants. To conclude, these clusters would be a good place to open our Mexican restaurant.